1 Glivenko-Cantelli type theorems

Given i.i.d. observations $X_1, ..., X_n$ with unknown distribution function $F(t)$, consider the empirical (sample) CDF

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{[X_i \leq t]}.$$

Then as $n \to \infty$,

$$\sup_{-\infty < \ t \ < \infty} |\hat{F}_n(t) - F(t)| \xrightarrow{a.s.} 0$$

Without the sup (i.e. for a fixed t) this is just an ordinary LLN for Bernoulli r.v.s The difficult (and usefulness) is in the sup. Notice that $F(t) = P(X \leq t) = P(X \in (-\infty, t])$, where $(-\infty, t]$ can be considered as a set A (indexed by t). And the Glivenko-Cantelli theorem can be rewritten as:

$$\sup_{A} | \int_A d[\hat{F}_n(s) - F(s)]| \xrightarrow{a.s.} 0$$

Does the following convergence hold if A is any Lebesgue measurable set in $\mathfrak{F}$?

$$\sup_{A \in \mathfrak{F}} | \int_A d[\hat{F}_n(s) - F(s)]| \xrightarrow{a.s.} 0$$

We know the following:
(1) if $\mathfrak{F} = \{(-\infty, t], \forall t \in R\}$, then the uniform convergence holds;
(1.5) if $\mathfrak{F} = \{(a, b]$, for any real $a < b\}$, then the uniform convergence holds;
(2) if $\mathfrak{F} = \{$ all measurable sets $\}$, then the uniform convergence doesn't holds;
(3) if $\mathfrak{F} =$ Vapnik-Chervonenkis (V-C) sets, then the uniform convergence holds.
We shall see that the key is $\mathfrak{F} \cap \{x_1, x_2, \cdots, x_n\}$ should have n^k (polynomials many) different sets, not exponentially many (2^n).

1.1 The proof of Glivenko-Cantelli theorem

Suppose $X_1, ..., X_n \overset{i.i.d.}{\sim} F(t)$, and $Y_1, ..., Y_n \overset{i.i.d.}{\sim} F(t)$ (same CDF). Also assume X's are independent of Y's. Let

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{[X_i \leq t]}$$

and

$$F_n^{\star}(t) = \frac{1}{n}\sum_{i=1}^{n} I_{[Y_i \leq t]}$$

Step 1: Symmetrization (See Page 14 of Pollard for details)

$$\forall \epsilon > 0; \qquad P(\sup_{-\infty < t < \infty} |\hat{F}_n(t) - F(t)| > \epsilon)$$

$$\leq 2P(\sup_{-\infty < t < \infty} |(\hat{F}_n(t) - F(t)) - (F_n^{\star}(t) - F(t))| > \frac{\epsilon}{2})$$

$$= 2P(\sup_{-\infty < t < \infty} |\hat{F}_n(t) - F_n^{\star}(t)| > \frac{\epsilon}{2})$$

Since $\hat{F}_n(t)$ and $F_n^\star(t)$ are piecewise constant functions, thus $|\hat{F}_n(t) - F_n^\star(t)|$ has at most $(2n+1)$ different values when $-\infty < t < \infty$.

Step 2: Turn infinite many "Sup" to finite many "Max", corresponding to (2n+1) different values.

$$
\begin{aligned}
2P(\sup_{-\infty<t<\infty} |\hat{F}_n(t) - F_n^\star(t)| > \frac{\epsilon}{2}) &= 2P(\max_{t=t_1,\dots,t_{2n+1}} |\hat{F}_n(t) - F_n^\star(t)| > \frac{\epsilon}{2}) \\
&= 2P(\bigcup_{i=1}^{2n+1} |\hat{F}_n(t_i) - F_n^\star(t_i)| > \frac{\epsilon}{2}) \\
&\le 2\sum_{i=1}^{2n+1} P(|\hat{F}_n(t_i) - F_n^\star(t_i)| > \frac{\epsilon}{2}) \qquad \text{(By Boole's ineq.)}
\end{aligned}
$$

Step 3: Hoeffding's Inequality (Pollard, 1984)

Suppose $Y_1{}^\star, ..., Y_n{}^\star$ are independent with $EY_i{}^\star = 0$ (Mean 0) and $a_i \le Y_1{}^\star \le b_i$ (bounded) then,

$$\forall \eta > 0, \quad P(|Y_1{}^\star + Y_2{}^\star + ... + Y_n{}^\star| > \eta) \le 2e^{\frac{-2\eta^2}{\sum_{i=1}^n (b_i - a_i)^2}}.$$

Let

$$Y_i{}^\star = \frac{1}{n}(I_{[X_i \le t]} - I_{[Y_i \le t]})$$

then we have

$$-\frac{1}{n} \le Y_i{}^\star \le \frac{1}{n}$$

and $E(Y_i{}^\star) = 0$. Thus Hoeffding's Inequality can be applied to $|\hat{F}_n(t_i) - F_n^\star(t_i)|$, with $\eta = \frac{\epsilon}{2}$

$$
\begin{aligned}
2\sum_{i=1}^{2n+1} P(|\hat{F}_n(t_i) - F_n^\star(t_i)| > \frac{\epsilon}{2}) &\le 2\sum_{i=1}^{2n+1} 2\exp\left(\frac{-2(\frac{\epsilon}{2})^2}{(\frac{2}{n})^2 n}\right) \\
&= (8n+4)e^{-\frac{n\epsilon^2}{8}} \\
&\to 0 \qquad \text{as } n \to \infty
\end{aligned}
$$

Remarks: (1) The above inequality holds for any $\epsilon > 0$ and any n. So we actually proved

$$P(\sup_{-\infty<t<\infty} |\hat{F}_n(t) - F(t)| > \epsilon) \le (8n+4)e^{-\frac{n\epsilon^2}{8}}; \tag{1}$$

(2) It is worth noting that how fast this bound $(8n+4)e^{-\frac{n\epsilon^2}{8}}$ goes to 0. For example, $\sum_{n=1}^\infty (8n+4)e^{-\frac{n\epsilon^2}{8}} < \infty$, an application of Borel-Cantalli lemma turns this into a.s. convergence. so Glivenko-Cantelli is almost surely converge. This also works if we replace $(8n+4)$ with any polynomials of n like n^k.

1.2 Generalizations

Many generalizations are possible.

1. The random variables $X_1, X_2, \cdots, X_n$ need only be independent; and do not have to be identically distributed. The limiting distribution is then $\bar{F}_n(t) = 1/n \sum F_i(t)$. (The limit is always obtained by replace the random variables by the expectations)

2. The constant $1/n$ may be replaced by other constants or a sequence of n constants: $a_1, a_2, \cdots, a_n$. The result will be

$$P(\sup_{-\infty<t<\infty} \sum_{i=1}^{n} |a_i I[X_i \le t] - a_i F_i(t)| > \epsilon) \le (8n+4) \exp\left[-\frac{\epsilon^2}{8 \sum_{i=1}^{n} 1/a_i^2} \right];$$

3. The limit do not have to be distribution functions. Any bounded non random function will do. In particular a sub-distrbution function.

$$\sup_t \sum_{i=1}^{n} a_i |I_{[X_i \le t,\ \delta_i = 1]} - U_i(t)|$$

where $U_i(t) = EI_{[X_i \le t,\ \delta_i=1]}$.

Excercise:

Suppose, as $n \to \infty$ we have

$$\sup_{-\infty<t<\infty} \frac{1}{n} |N(t) - EN(t)| \longrightarrow^{a.s.} 0 \quad \text{and} \quad \sup_{-\infty<t<\infty} \frac{1}{n} |R(t) - ER(t)| \longrightarrow^{a.s.} 0$$

as $n \to \infty$. Show that

$$\int_0^t \frac{dN(s)}{R(s)} \longrightarrow \int_0^t \frac{dEN(s)}{ER(s)}$$

again, uniformly for those t that $ER(t) > \eta > 0$.

Furthermore, suppose $g(t)$ is a function that the integral in the limit below is well defined. Let $g_n(t)$ be a random sequence of functions that

$$\sup_{-\infty<t<\infty} |g_n(t) - g(t)| \to^{a.s.} 0 \ .$$

Show that

$$\int_0^t \frac{g_n(s) dN(s)}{R(s)} \longrightarrow \int_0^t \frac{g(s) dEN(s)}{ER(s)}$$

again, uniformly for those t that $ER(t) > \eta > 0$.

Let

$$\mathfrak{F} = \{A_t = (-\infty, t], -\infty < t < \infty\}$$
$$A_t \cap \{x_1, ..., x_n\} = \{\phi\}, \{x_1\}, \{x_1x_2\}, ..., \{x_1...x_n\}$$

(WLOG assume the x_i's are ordered.) The number of *all* subsets of $\{x_1, ..., x_n\}$ is 2^n, but the number of all sets of the form $A_t \cap \{x_1, ..., x_n\}$ is (n + 1). In general, if the number of all sets of the type $A \cap \{x_1, ..., x_n\}$ is a polynomial function in n (i.e. $O(n^k) \ll 2^n$), then the sets contained in A is a V-C class of sets.
For example, if $A = A_{ab} = (a, b], -\infty < a < b < \infty$, then the number of all sets of type $A \cap \{x_1, ..., x_n\}$ is $\frac{n(n+1)}{2} + 1$ (including empty set). Therefore the sets of $A_{ab} = (a, b]$ is a V-C class of sets.
Claim: If and only if $\mathfrak{F}$ is a V-C class of sets, then

$$P(\sup_{A \in \mathfrak{F}} |\int I_{[A]} d\hat{F}_n(t) - \int I_{[A]} dF(t)| > \epsilon) \to 0$$

1.3 Applications

In the Cox model, the Breslow estimate of Baseline hazard and Fisher information matrix.

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\frac{1}{n} dN(s)}{\frac{1}{n}\sum_{i=1}^n I_{[Y_i \geq s]} e^{\beta z_i}}$$

We focus on the denominator.

$$P\left(\sup_s |\frac{1}{n}\sum_{i=1}^n I_{[Y_i \geq s]} e^{\beta z_i} - \frac{1}{n}\sum_{i=1}^n P(Y_i \geq s) e^{\beta z_i}| > \epsilon\right)$$
$$\leq (8n+4) e^{-\frac{n\epsilon^2}{8M^2}} \quad (Condition : |z_i| \leq M < \infty)$$

Where

$$P(Y_i \geq s) = P(T_i \geq s) P(C_i \geq s) = e^{\Lambda_i(s)}[1 - G(s)] = e^{\Lambda_0(s) e^{\beta z_i}}[1 - G(s)]$$

Similar for Fisher information matrix

$$\frac{1}{n}\sum_{i=1}^n {z_i}^2 I_{[Y_i \geq s]} e^{\beta z_i}$$

The Glivenko-Cantelli can also be formulated for functions.

$$P(\sup_{f\in\mathfrak{F}} |\int f(x)d\hat{F}_n(x) - \int f(x)dF(x)| > \epsilon)$$

$$= P(\sup_{f\in\mathfrak{F}} |\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int f(x)dF(x)| > \epsilon)$$

What is the condition on $\mathfrak{F}$ to make above $\to 0$?
V-C class of function: if a function's graph is a V-C class of sets.

$$f(x) \Longleftrightarrow graph\{(x,y)|f(x) > y\}$$

More dimensions (example: 2 dimensions)
The number of all sets of $A \cap \{x_1, ..., x_n\}$ is a polynomial function in $n \Rightarrow A = rectangles \in$ "V-C class of sets"
Hence the Glivenko-Cantelli convergence works in 2 dimensions etc.

Homework:
Is the following true? Prove if it is true.

$$\sup_{-\infty<t\leq x_{(n)}} \left| \frac{1}{1-\hat{F}_n(t)} - \frac{1}{1-F(t)} \right| \overset{a.s.}{\to} 0$$

If not, what bound instead of $x_{(n)}$ will make the convergence hold?

Homework: Suppose $\hat{\Lambda}_n(t)$ is the Nelson-Aalen estimator based on n right censored observations, and the $\Lambda(t)$ is the true cumulative hazard. Assume $\Lambda(t)$ is continuous, also assume $\Lambda(t) \uparrow \infty$ as $t \uparrow \infty$. Show that, as $n \to \infty$

$$\sup_{t\leq M} |\hat{\Lambda}_n(t) - \Lambda(t)| \longrightarrow 0$$

either in probability or almost surely.

Any speed? Can we make it $\sup_{t<\infty}$?

Reference: Pollard D. (1984) Convergence of Stochastic Processes. Springer

2 Empirical Likelihood and Bootstrap

The idea of Boostrap: In the correspondence (or the link between that) of $\hat{F}_n(\cdot) \longrightarrow (\hat{\theta}_n - \theta_0)$, bootstrap apply a random perturb to the $\hat{F}_n$, and see how $(\hat{\theta}_n - \theta_0)$ change accordingly. Repeat this many times and you have a sampling distribution of $(\hat{\theta}_n - \theta_0)$. The random perturbation is obtained by a random sampling (or re-sampling) to $\hat{F}_n$.

The idea of Empirical Likelihood: In the correspondence of $\hat{F}_n(\cdot) \longrightarrow \hat{\theta}_n$, EL force the statistic $\hat{\theta}_n$ to the value θ_0, and find the tilted $\hat{F}_n$ that corresponding to this perturbed $\hat{\theta}_n$. We denote the tilted distribution as $\hat{F}_n^\lambda$ for some nonzero λ.

It turns out

$$-2 \log \frac{EL(\hat{F}_n^\lambda)}{EL(\hat{F}_n)}$$

will have a chi square distribution, a pivatol distribution when θ_0 is the true value of the parameter.

Under null hypothesis, the perturbation of $\hat{\theta}_n$ to θ_0 is of order $1/\sqrt{n}$ (usually). In bootstrap, the perturbation of re-sampling to $\hat{F}_n$ is also of order $1/\sqrt{n}$.

The differrence: the bootstrap is a random perturbation but EL is a fixed perturbation, so bootstrap usually need simulation to repeat many times, and result may be slightly difference due to random re-sample errors. On the other hand, bootstrap can be applied to any statistic, but EL works most successfully for the case $\hat{\theta}$ is NPMLE. (has anyone try it on non-MLE?) In some setup, it may not be clear how a random perturbation should be applied to the $\hat{F}$ becausse there are several plausible ways to do it. On the other hand, for EL there is usually clear, and only one way to set $\hat{\theta}$ to θ_0.

Bootstrap needs to estimate a whole distribution (or percentile), and the EL can rely on the fact that the distribution of the likelihood ratio is a pivatol chi square.

The introduction of the λ turns the non-parametric problem into a parametric problem. In the new parametric problem, we are estimate the "true" value of zero, and the information of λ is just the negative second derivative of the log likelihood and the MLE is $\hat{\lambda}_n$ which has an asymptotic normal distribution. This sub-family of parametric distributions are the so-called least favorable sub-distributions, an idea first proposed by Stone in 1956.

For any (square) integrable function $\phi(t)$ and a distribution $F(\cdot)$, define

$$\bar{\phi}(t) = \bar{\phi}_F(t) = \frac{1}{1 - F(t)} \int_{(t,\ \tau_F]} \phi(u) dF(u)$$

where $\tau_F = \sup\{x : F(x) < 1\}$.

Theorem Denote the Kaplan-Meier estimator based on n i.i.d. observations as $\hat{F}_n$. We have

$$\frac{1}{1 - \hat{F}_n(t)} \int_{(t,\ \tau_{\hat{F}_n}]} \phi(u) d\hat{F}_n(u) \longrightarrow \frac{1}{1 - F(t)} \int_{(t,\ \tau_F]} \phi(u) dF(u)$$

that is

$$\bar{\phi}_{\hat{F}_n}(t) \longrightarrow \bar{\phi}_F(t)$$

The convergence is uniformly, almost sure, i.e.

$$\sup_t |\bar{\phi}_{\hat{F}_n}(t) - \bar{\phi}_F(t)| \longrightarrow 0, \quad a.s.$$

Theorem Assume $\phi(t)$ is square integrable with respect to $F(t)$. Then we have

$$\int [\phi(t) - \bar{\phi}_{\hat{F}}(t)]^2 d\hat{F}_n(t) \longrightarrow \int [\phi(t) - \bar{\phi}_F(t)]^2 dF(t)$$

Akritas (2000) studied the central limit theorem for the Kaplan-Meier integrals. There are earlier papers about the same topic, but the asymptotic variance expression of Akritas (2000) is new and interesting.

Theorem (Akritas 2000) The asymtotic variance of Kaplan-Meier integrals are

$$AsyVar\left(\sqrt{n}\int \phi(t)d\hat{F}_{KM}(t)\right) = \int_{-\infty}^{\tau}[\phi(t)-\bar{\phi}(t)]^2\frac{[1-F(t)]dF(t)}{1-H(t-)} \ .$$

A multivariate version of this theorem can be easily obtained. Denote $\Phi(t) = (\phi_1(t), \cdots, \phi_k(t))$, then the asymptotic variance-covariance matrix of the k-vector of Kaplan-Meier integrals is

$$AsyVarCov\left(\sqrt{n}\int \Phi(t)d\hat{F}_{KM}(t)\right) = [\sigma_{ij}] \ ,$$

with

$$\sigma_{ij} = \int_{-\infty}^{\tau}[\phi_i(t)-\bar{\phi}_i(t)][\phi_j(t)-\bar{\phi}_j(t)]\frac{[1-F(t)]dF(t)}{1-H(t-)} \ .$$

This multivariate version can be obtained by using the representation of Akritas (2000), his Theorem 6.

An easier to check sufficient condition to insure the variance are well defined is

$$\int_{-\infty}^{\tau}\frac{\phi^2(s)}{1-G(s-)}dF(s) < \infty \ .$$

When there is no censoring, the Kaplan-Meier estimator become the empirical distribution and the integral with respect to empirical distribution is just the i.i.d. summation (or average). Finally, when there is no censoring $1-H(s-) = 1-F(s-)$, the covariance formula of Akritas above simplify to the following

$$AsyCov\left(\frac{1}{\sqrt{n}}\sum_{u=1}^{n}\phi_i(X_u), \ \frac{1}{\sqrt{n}}\sum_{u=1}^{n}\phi_j(X_u)\right)$$

can be written as

$$\int_{-\infty}^{\tau}[\phi_i(t)-\bar{\phi}_i(t)][\phi_j(t)-\bar{\phi}_j(t)]\frac{[1-F(t)]dF(t)}{1-F(t-)} \ .$$

On the other hand, the said covariance can obviously be written as

$$\int_{-\infty}^{\tau}[\phi_i(t)-E\phi_i][\phi_j(t)-E\phi_j]dF(t) \ .$$

We, therefore, arrive at the following identity

Lemma For function ϕ_i and ϕ_j that are square integrable with respect to $F(t)$ we have

$$\int_{-\infty}^{\tau}[\phi_i(t)-\bar{\phi}_i(t)][\phi_j(t)-\bar{\phi}_j(t)]\frac{[1-F(t)]dF(t)}{1-F(t-)} = \int_{-\infty}^{\tau}[\phi_i(t)-E\phi_i][\phi_j(t)-E\phi_j]dF(t) \ .$$

When either the expectations $E\phi_i = 0$ or $E\phi_j = 0$ or both, the above identity can further be simplified to

$$\int_{-\infty}^{\tau}[\phi_i(t)-\bar{\phi}_i(t)][\phi_j(t)-\bar{\phi}_j(t)]\frac{[1-F(t)]dF(t)}{1-F(t-)} = \int_{-\infty}^{\tau}[\phi_i(t)][\phi_j(t)]dF(t) \ .$$

We comment that this identity holds for *any* distribution $F(\cdot)$, we later will use this when $F(t)$ is the Kaplan-Meier distribution.

A general empirical likelihood theorem. For a sample of n independent observations with distribution belongs to a family $F_n(\beta)$ here β is the finite dimentional parameter, F_n can be nonparametric. If there exist a distribution F_{0n} such that $F_n << F_{0n}$, that is all distributions are dominated by a single (but can depend on n) distribution F_{0n}, then empirical likelihood works for test the finite parameter β of the distributions F_n.